UANL

UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN

FCFM

FACULTAD DE CIENCIAS FÍSICO MATEMÁTICAS

Universidad Autónoma de Nuevo León

Facultad de Ciencias Físico Matemáticas

# Producto Integrador de Aprendizaje

## Aprendizaje automático

Nombre: Sergio Andrés Elizondo Rodríguez

Grupo: 001

Maestro: José Anastacio Hernández Saldaña

Viernes 26 de Julio del 2024

# Reporte de resultados

**Objetivo:** Encontrar el mejor modelo de clasificación utilizando validación cruzada y el criterio de área bajo la curva característica operativa del receptor (*ROC AUC*), logrando un valor mayor a 0.75 en el conjunto de validación y prueba.

**Datos utilizados:**

- **Conjunto de entrenamiento:** 9239 observaciones, 10 columnas.
- **Conjunto de prueba:** 2309 observaciones, 9 columnas.

## Procesamiento previo de los datos

- Se convirtió la variable objetivo (*engagement*) a valores numéricos.
- Se separaron las características (*X*) y las etiquetas (*y*) en el conjunto de entrenamiento.
- Se normalizaron las características utilizando un escalador (*standard scaler*).
- Se dividió el conjunto etiquetado (*entrenamiento*) en datos de entrenamiento y de validación (80% y 20%, respectivamente).

## Selección de modelo

Se evaluaron los siguientes modelos:

- Regresión logística
- *K* vecinos más cercanos (*KNN*)
- Clasificador de margen máximo (*support vector classifier*)
- Árbol de decisión
- Bosque aleatorio (*random forest*)
- Potencialización de gradiente (*gradient boosting*)
- *XGBoost*
- *LightGBM*

Para la evaluación, se definieron espacios de búsqueda de hiperparámetros específicos para cada modelo, optimizando su rendimiento mediante validación cruzada de 5 pliegues en una búsqueda en cuadrícula (*grid search cross-validation*) con el criterio *ROC AUC*.

# Resultados

A continuación se presentan los resultados de la métrica *ROC AUC* obtenidos en el conjunto de validación para los mejores hiperparámetros de cada modelo:

| **Modelo** | **Mejores hiperparámetros** | **Área bajo la curva operativa del receptor (*ROC AUC*)** |
|---|---|---|
| **Árbol de decisión** | • **Ajuste de pesos:** Ninguno<br>• **Criterio de división:** Entropía<br>• **Profundidad máxima:** Ilimitada<br>• **Máximo de características por división:** 80%<br>• **Máximo de nodos hoja:** 40<br>• **Umbral para reducción de impureza:** 0<br>• **Mínimo de muestras por hoja:** 5<br>• **Mínimo de muestras por división:** 5<br>• **Fracción mínima del peso por hoja:** 0.01<br>• **Estrategia de división:** Elegir la mejor | 85.93% |
| *K* vecinos más cercanos (*KNN*) | • **Vecinos a considerar:** 55<br>• **Potencia de la distancia de Minkoswki:** 2 (distancia euclidiana)<br>• **Ponderación:** Por distancia | 87.68% |
| Regresión logística | • **Parámetro de regularización inversa:** 0.02<br>• **Ajuste de pesos:** Balanceado<br>• **Tipo de regularización:** L2 (*ridge*)<br>• **Algoritmo de optimización:** *liblinear* | 87.84% |
| Clasificador de margen máximo (*support vector classifier*) | • **Parámetro de regularización:** 3<br>• **Núcleo (*kernel*):** Polinómico<br>• **Constante del núcleo:** 19<br>• **Grado del polinomio:** 3<br>• **Coeficiente γ:** 0.04 | 89.07% |
| Bosque aleatorio (*random forest*) | • **Ajuste de pesos:** Ninguno<br>• **Criterio de división:** Entropía<br>• **Profundidad máxima:** 10<br>• **Máximo de características por división:** 75%<br>• **Mínimo de muestras por hoja:** 1<br>• **Mínimo de muestras por división:** 2<br>• **Número de árboles:** 200 | 89.25% |
| *XGBoost* | • **Amplificador (*booster*):** Regresión múltiple aditiva (*DART*)<br>• **Reducción mínima (γ):** 0.1<br>• **Tasa de aprendizaje:** 0.05<br>• **Profundidad máxima:** 5<br>• **Peso mínimo por hoja:** 2<br>• **Restricciones monotónicas:** (1, 0)<br>• **Número de árboles:** 97 | 90.31% |

| | | |
|---|---|---|
| | • **Función de pérdida:** Logística binaria<br>• **Término de regularización L1 (*LASSO*):** 1<br>• **Término de regularización L2 (*ridge*):** 0<br>• **Peso de las clases positivas:** 2<br>• **Fracción de muestras para entrenamiento:** 90% | |
| *LightGBM* | • **Tasa de aprendizaje:** 0.01<br>• **Profundidad máxima:** Ilimitada<br>• **Mínimo de muestras por hoja:** 30<br>• **Número de árboles:** 350<br>• **Máximo de hojas por árbol:** 40 | 90.31% |
| Potencialización de gradiente (*gradient boosting*) | • **Tasa de aprendizaje:** 0.1<br>• **Función de pérdida:** Exponencial<br>• **Profundidad máxima:** 5<br>• **Mínimo de muestras por hoja:** 1<br>• **Mínimo de muestras por división:** 2<br>• **Fracción mínima del peso por hoja:** 0.03<br>• **Número de árboles:** 100 | 90.45% |

El modelo de **potencialización de gradiente (*gradient boosting*)** fue el que obtuvo el mejor rendimiento (*ROC AUC* = 90.45%).

## Evaluación

Para la evaluación final, se utilizó el mejor modelo encontrado (*gradient boosting*) para predecir las probabilidades en el conjunto de prueba. Las predicciones se guardaron en el archivo “resultados.csv”, asignando la probabilidad de participación (*engagement*) predicha para cada identificador (*id*).

## Conclusión

El modelo de potencialización de gradiente (*gradient boosting*) demostró ser el más efectivo en la evaluación de área bajo al curva operativa del receptor (*ROC AUC*), con una métrica de 0.9045, superando el umbral de 0.75 en el conjunto de validación, lo cual confirma que el modelo seleccionado es adecuado para la tarea de clasificación planteada. Este rendimiento superior puede atribuirse a varios factores, como la optimización de una función de pérdida robusta frente a valores atípicos y la consideración de un gran número de árboles de decisión, lo cual permite que los errores de un árbol se minimicen por los demás árboles, permitiendo a su vez capturar relaciones no lineales entre los datos.

Además el ajuste fino de hiperparámetros permitió mejorar la capacidad del modelo de capturar patrones específicos de nuestro conjunto de datos sin caer en el sobreajuste (gracias a la validación cruzada). Es necesario contar con las etiquetas en el conjunto de prueba para evaluar si efectivamente el bueno comportamiento del modelo se sigue replicando con los mismos.